Les grilles pour le développement médical

N. Jacq, Matthieu Matthieu.Reichstadt@inrae.Fr Reichstadt, F. Jacq, J. Salzemann, M. Zimmermann, A. Maas, M. Sridhar, K. Vinodkusam, H. Schwichtenberg, M. Hofmann, et al.

Les grilles pour le développement médical

N. Jacq[1+2], M. Reichstadt[1], F. Jacq[1], J. Salzemann[1], M. Zimmermann[3], A. Maas[3], M. Sridhar[3], K. Vinodkusam[3], H. Schwichtenberg[3], M. Hofmann[3], V. Breton[1]

[1]Laboratoire de Physique Corpusculaire, Université Blaise Pascal/IN2P3-CNRS UMR 6533, France
[2] Communication & Systèmes, CS-SI, France
[3] Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Department of Bioinformatics, GERMANY

Résumé

Le développement récent des sciences et technologies de l'information et de la communication permet aujourd'hui la création de véritables infrastructures pour le calcul et le stockage de données hétérogènes à l'échelle régionale, nationale et internationale. Ces infrastructures, appelées grilles informatiques, permettront bientôt d'utiliser les ressources informatiques mutualisées avec autant de facilité que nous utilisons aujourd'hui l'électricité.

L'utilisation des grilles afin d'accélérer la découverte de médicaments est une voie très prometteuse pour l'avenir. Par cette approche *in silico*, le nombre de molécules ainsi que la vitesse de test peuvent être grandement augmentés induisant un coût moindre de développement de médicaments.

Du 11 Juillet au 31 Août 2005, l'expérience WISDOM (Wide *In Silico* Docking On Malaria) a permis de tester rien moins qu'un million de ligands (médicaments potentiels) pour le traitement du paludisme: 1700 ordinateurs à travers le monde ont ainsi été associés à cette démarche permettant de réaliser en un mois ce qui aurait nécessité 80 ans sur un ordinateur classique. L'analyse des résultats est en cours.

Par cette approche, on peut souhaiter également que les maladies orphelines puissent bénéficier d'un intérêt nouveau de la part des industries pharmaceutiques, à travers notamment la baisse du coût de développement d'un médicament, principal obstacle actuellement à leur mobilisation.

Mots clefs

Grille informatique, maladie négligée, *in silico* docking, paludisme

I. Introduction

Avec les formidables progrès de la génomique, de la bioinformatique et de l'informatique médicale, les centres hospitaliers et les laboratoires de recherche se retrouvent confrontés à la nécessité d'archiver, de gérer et d'analyser des volumes de données considérables.

Dans ce contexte, la « santé » n'implique plus seulement la notion de la pratique clinique mais couvre tous les domaines scientifiques depuis l'information au niveau moléculaire (génomique, post-génomique, protéomique, …) à travers les cellules et les tissus jusqu'à l'individu pour finalement atteindre le niveau macroscopique des populations (santé publique).

Développement récent des sciences et technologies de l'information et de la communication, les grilles informatiques permettent aujourd'hui la création de véritables infrastructures pour le calcul et le stockage de données hétérogènes à l'échelle régionale, nationale et internationale. Elles constituent des environnements privilégiés pour stocker, analyser et gérer l'ensemble des informations biologiques et médicales nécessaires à une médecine de plus en

plus personnalisée. De nombreuses initiatives voient le jour en Europe pour améliorer la recherche en biologie moléculaire et en médecine grâce aux grilles.

Parmi ces initiatives, plusieurs se concentrent sur l'utilisation des grilles informatiques pour la recherche sur les maladies négligées. Ainsi, plusieurs projets ont vu le jour en Europe (WISDOM [1], Africa@Home [2], Dengue grid [3]) pour accélérer la recherche de nouveaux médicaments *in silico* (c'est-à-dire à l'aide d'un ordinateur) sur grille. Dans cet article, nous allons décrire le principe d'une grille informatique et son champ d'application aux maladies négligées. Puis, nous présenterons le projet WISDOM de recherche de nouveaux médicaments contre le paludisme.

II. Les grilles informatiques

Les grilles informatiques permettront bientôt d'utiliser les ressources informatiques mutualisées avec autant de facilité que nous utilisons aujourd'hui l'électricité. Un utilisateur, qui s'abonne à une grille, pourra alors exécuter des calculs ou stocker des informations sans avoir à définir les machines qu'il va mobiliser. Cette technologie permet, par exemple, de mieux tirer parti du potentiel d'un parc informatique. Un chercheur voulant mener une opération trop complexe sur sa machine pourra voir son calcul effectué dans un autre centre de recherche, où les machines sont momentanément sous-exploitées. De même dans le cas du stockage de ses données, un chercheur pourra faire appel à l'ordinateur d'un tiers à travers la grille et pourra aller consulter ces données par la suite. Le tout est piloté par un logiciel arbitre chargé de veiller à l'utilisation optimale et à la sécurité des ressources disponibles sur le réseau.
Les grilles informatiques sont similaires à des lignes de métro car elles transportent l'information entre différents sites exactement comme les métros transportent des passagers.
Elles sont différentes d'Internet en ce sens qu'Internet fournit de l'information statique mais ne gère pas les flux d'information.

La région Auvergne joue un rôle pionnier dans le développement de ces grilles, notamment pour les besoins spécifiques de la recherche dans le domaine de la santé et des sciences du vivant. AuverGrid [4], grille régionale en Auvergne, vise à fournir une infrastructure de ressources informatiques pour aider le développement de tous les acteurs régionaux de la vie économique. Cette infrastructure est composée de « grappes » de PCs et d'espaces de stockage répartis dans la région Auvergne dans les laboratoires de recherche, les parcs technologiques et les structures gouvernementales. AuverGrid est la première infrastructure de grille déployée au niveau régional en France qui rassemble des partenaires académiques et industriels.

En termes de calculs à grande échelle, la valeur ajoutée de la grille vient du partage de ressources distribuées car les équipes de recherche aussi bien que les PME sont confrontées à des besoins émergents de modélisation complexe des systèmes. De même, la gestion de la connaissance à travers l'accessibilité aux données et aux outils ouvre de nouvelles perspectives pour le développement de la recherche au niveau international. Ceci est particulièrement vrai dans le domaine des sciences de la vie et la recherche médicale où la gestion des données et l'interopérabilité sont une clé pour l'innovation médicale. Un autre avantage majeur des grilles de calcul est de favoriser les collaborations entre sites distants par le partage de l'information, des ressources et des services à travers les organisations virtuelles.

III. Les grilles pour les maladies négligées

La technologie de grille fournit l'environnement collaboratif pour permettre le couplage entre la recherche de biologie moléculaire et les activités sur le terrain. Elle propose un nouveau paradigme pour la collecte et l'analyse des informations distribuées car elle permet de ne plus centraliser l'ensemble des informations en un seul lieu.

Sur une grille, les données peuvent être stockées n'importe où et cependant être lues par un utilisateur de façon transparente. Les ressources de calcul de la grille sont aussi partagées et peuvent être mobilisées à la demande pour permettre des analyses comparatives ou du criblage virtuel à grande échelle.

La perspective ainsi ouverte est d'accroître la capacité de l'industrie pharmaceutique et des institutions de recherche publique de partager des informations distribuées diverses et complexes sur une maladie pour une exploration commune et un bénéfice mutuel. Le but est de faciliter des interactions significatives pour produire des médicaments et des insecticides moins chers pour les maladies des pays en voie de développement et accroître le retour sur investissement pour les nouveaux médicaments dans les pays en voie de développement.

La recherche sur les maladies négligées pourrait bénéficier grandement du déploiement des grilles informatiques à plusieurs niveaux.

Un accès partagé à des outils et des bases de données bioinformatiques

La recherche de nouvelles cibles thérapeutiques utilise abondamment des outils bioinformatiques : algorithmes, bases de données biologiques. Les grilles devraient permettre de démocratiser l'accès à ces outils bioinformatiques et notamment à des bases de données mises à jour. Ceci est particulièrement important pour les laboratoires des pays en développement qui ont besoin de disposer de ces outils pour leurs recherches.

Une accélération de la recherche in silico *de nouveaux médicaments*

Le principe de la recherche *in silico* de nouveaux médicaments est de rechercher des médicaments potentiels en calculant des probabilités d'ancrage de molécules. La simulation *in silico* requiert en entrée la structure de la protéine cible, provenant de cristallographie ou de modélisation par homologie, et une bibliothèque de ligands maintenus, filtrés et formatés. Pour un projet public de criblage virtuel de médicaments, une collection de ligands virtuels doit être construite à partir des bases de ligands existantes (PDB Ligand Chemistry [5], KEGG-Ligand [6], bibliothèques privées ou publiques…). Le docking *in silico* nécessite aussi de choisir un logiciel validé. Seuls quelques logiciels de docking sont disponibles librement dans le domaine public (i.e. Dock [7]). La grille devrait permettre l'utilisation de tels logiciels demandeurs en temps de calcul avec de grandes bases de ligands.

Une fédération de bases de données pour des tests cliniques dans des secteurs infectés

Les technologies de grilles ouvrent de nouvelles perspectives pour la préparation et le suivi de missions médicales dans les pays en voie de développement, mais aussi dans l'aide aux centres médicaux locaux en terme de télé consulting, télé diagnostique, suivi des patients et télé-enseignement (e-learning). Dans chaque hôpital, les données patients sont enregistrées

dans des bases de données. Une fédération de ces bases de données permettrait aux données médicales d'être conservées dans les hôpitaux derrière un pare-feu (firewall) tout en étant visibles, à travers un réseau sécurisé, de médecins ou de chercheurs en fonction des autorisations.
De telles fédérations de bases de données peuvent être utilisées dans un but d'épidémiologie aussi bien que de tests cliniques.

Un espace de connaissance sur une maladie

L'objectif ici est de rendre disponible toutes les informations pertinentes sur une maladie à toutes les parties intéressées. Le concept d'espace de connaissance est d'organiser l'information de manière à ce qu'elle soit accessible en quelques clics. Ce concept est déjà utilisé avec succès en interne par les laboratoires pharmaceutiques pour stocker la connaissance. La grille permet de construire un espace de connaissance distribué où chaque participant peut garder sa propre information dans son ordinateur local. Un ensemble de services de grille est proposé pour rendre l'information facilement consultable pour les différents clients (médecins, chercheurs…). Ces services prendraient avantages des développements réalisés dans le domaine de l'analyse sémantique de texte et de l'exploitation de textes pour l'extraction de l'information en biologie et dans la recherche génomique.

Surveillance des travaux de terrain pour contrôler le paludisme

Les outils de surveillance pour contrôler le paludisme dans les secteurs infectés incluent la réduction des contacts entre les moustiques infectés et les humains en éliminant les sites de reproduction, les larvicides, la pulvérisation des maisons par des insecticides, l'imprégnation des moustiquaires par des insecticides. L'application réelle de ces mesures réduirait la morbidité et la mortalité dues au paludisme. Cependant, étant donné le pourcentage très élevé de transmission de P. *falciparum*, et particulièrement quand le moustique A. *Gambiae* est prédominant, l'impact de ces outils peuvent être limités par la capacité des moustiques à développer des résistances, et par le besoin de maintenance de ces interventions pendant plusieurs années.
La surveillance à long terme de ces outils bénéficierait de technologies permettant de meilleures collectes et analyses de ces données distribuées dans les pays en voie de développement. En effet, Internet est maintenant accessible dans le monde entier. L'idée est donc d'organiser la collecte d'information autour de centres agissant comme des répertoires locaux. Ces répertoires seraient fédérés au niveau mondial grâce à la technologie de grille pour des agences internationales ou des organisations à but non lucratif dans le but d'obtenir une vue générale des travaux réalisés dans le contrôle du vecteur.

IV. *In silico* docking dans un environnement de grille, l'expérience WISDOM

Introduction

Les avancées en chimie combinatoire permettent aujourd'hui de synthétiser des millions de molécules chimiques différentes. Ainsi ces millions de composés chimiques sont disponibles dans les laboratoires et également dans les bases de données électroniques 2D et 3D. Mais il est presque impossible de filtrer un tel nombre de composés dans des laboratoires expérimentaux par criblage haut débit. Au-delà du coût très élevé, la proportion de molécules intéressantes est très faible. Il est de l'ordre de 1 pour 100 000 composés quand le criblage est réalisé sur des cibles comme les enzymes [8].

Une alternative est le criblage haut débit virtuel par docking moléculaire, ou amarrage moléculaire, une technique qui permet de cribler des millions de composés rapidement, efficacement et à moindre coût. Cribler des millions de composés chimiques *in silico* (c'est-à-dire sur ordinateur) est un processus complexe requérant une gestion fine des données et des tâches de calcul. Cribler chaque composé prend de quelques minutes à plusieurs heures sur un ordinateur classique, en fonction de la complexité structurale de la molécule. Cela signifie que cribler tous les composés d'une base de données pourrait durer des années. Ce temps de calcul peut être réduit significativement avec une grille informatique rassemblant des milliers d'ordinateurs [9].

WISDOM (Wide *In Silico* Docking On Malaria) est une initiative pour le déploiement d'une application de docking à grande échelle contre le paludisme. Il s'agit d'une première étape pour mettre en place une recherche de médicaments *in silico* sur une infrastructure de grille. 3 objectifs motivent l'initiative. L'objectif biologique est de proposer de nouveaux inhibiteurs pour une famille de protéine produite par plasmodium falciparum. L'objectif en chimie informatique est de déployer une application de docking sur une infrastructure de grille. L'objectif en informatique de grille est le déploiement d'une application très demandeuse en temps de calcul et générant une grande quantité de données pour tester l'infrastructure de grille et ses services. Un tel déploiement est appelé un challenge de production de données, ou data challenge.

Objectifs et composants de l'expérience WISDOM

Le paludisme est une maladie infectieuse touchant 300 millions de personnes et tuant 1,5 million de personnes chaque année [10]. Le paludisme est causé par un parasite protozoaire, le plasmodium. Il y a plusieurs médicaments anti-paludique disponibles aujourd'hui. Mais l'émergence constante de résistances et le coût des médicaments actuels montrent l'importance d'explorer de nouvelles stratégies pour combattre le paludisme [10]. L'une de ces stratégies concerne le métabolisme de l'hémoglobine, qui est l'une des clefs de la survie du parasite. Plusieurs protéases du plasmodium sont impliquées dans la dégradation de l'hémoglobine humaine à l'intérieur de la vacuole du parasite, dans les érythrocytes. La plasmepsine, la protéase aspartique de plasmodium, est responsable du clivage initial de l'hémoglobine humaine [11]. Il y a 10 plasmepsines différentes codées par 10 gènes différents de *P. falciparum* [12]. Ces plasmepsines ont un haut niveau d'homologie de séquences (65-70%). Elles partagent en même temps seulement 35% d'homologie avec la protéase aspartique humaine la plus proche, la Cathepsine D4 [13]. Ces arguments, en plus de la présence de données cristallographiques X de bonne qualité, font de la plasmepsine une cible particulièrement intéressante pour la conception de médicaments dans le cadre de l'expérience WISDOM.

Le docking est l'une des premières étapes dans la conception de médicaments. Le docking protéine-composé chimique est la prédiction de l'énergie de liaison entre une protéine cible et une base de données de composés chimiques, ou ligands. L'objectif est d'identifier quelles molécules peuvent se lier sur le site actif d'une protéine dans le but d'inhiber son action et donc d'interférer avec un processus moléculaire essentiel pour le pathogène. Les bibliothèques des structures 3D des ligands sont accessibles gratuitement par les compagnies chimiques qui peuvent les produire. Ainsi 2 logiciels de docking ont été choisis pour leur méthode algorithmique internationalement reconnues : Autodock [14], logiciel librement disponible pour la recherche publique, et FlexX [15], logiciel commercial mis gracieusement à la disposition de l'expérience WISDOM par l'entreprise BioSolvIT [16]. 1 million de composés chimiques étudiés ont été récupérés à partir de la base de données ZINC [17]. De

larges ressources sont nécessaires pour tester une famille de protéines avec un grand nombre de médicaments candidats, plusieurs jeux de paramètres et des logiciels de docking.

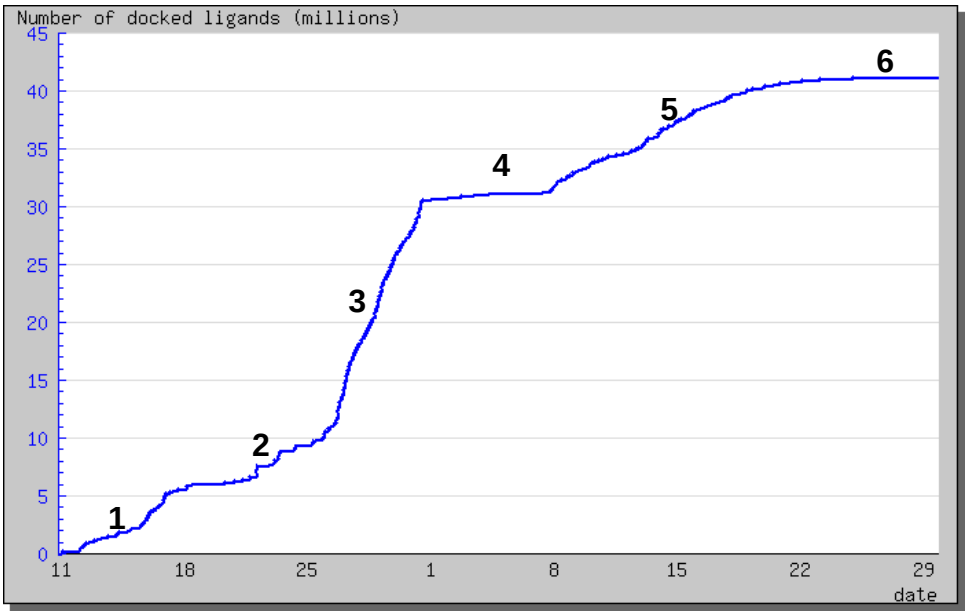
Le docking est un type d'application facilement distribuable sur une grille. Chaque tâche de comparaison d'une protéine avec un ligand est indépendante, et un ensemble de ces tâches peut être envoyé sur un nœud de calcul de la grille. Les données stockées sur les éléments de stockage sont alors transférés sur le nœud de calcul, puis les résultats sont stockés sur un élément de stockage de la grille, et répliqués sur d'autres éléments pour réaliser une copie de sauvegarde. De nombreuses ressources de calcul et de stockage ont été mises à disposition par le projet Enabling Grids for E-sciencE (EGEE) [18]. Le projet EGEE est financé par la Commission européenne et a pour but de construire sur les plus récentes avancées des technologies de grille et de développer un service d'infrastructure de grille disponible 24h sur 24. Les ressources mises à disposition pour les applications biomédicales sont représentées sur la carte ci-dessous.

Même si de nombreuses applications sont aujourd'hui déployées sur grille, seulement quelques unes le sont à une très grande échelle [19], et aucune dans le domaine biomédical. C'est pourquoi l'expérience WISDOM a été l'occasion de tester réellement l'efficacité de la grille et de ses services.

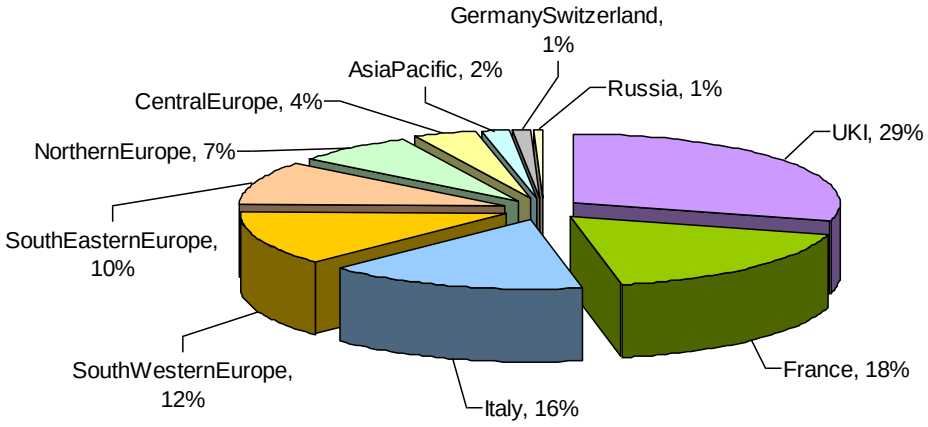
Le déploiement de WISDOM

Les différents éléments de l'application ont été installés sur la grille EGEE. Un environnement spécifique de définition des tâches de calcul, de soumission sur la grille, de gestion des états de ces tâches, puis de récupération des résultats, a été développé pour l'expérience. Un serveur de licence a également été mis en place pour le logiciel commercial FlexX dont l'accès a ainsi été restreint pendant les 3 semaines de production avec cet outil.

Plus de 40 millions de complexes protéine-ligand ont été étudié en 6 semaines environ. La figure suivante montre la production de ces complexes au cours du temps.



Plusieurs phases de production intensive (1, 3, 5), de re-soumissions de tâches manquantes (2, 6), et de pause (4) sont visibles.
80 années de calcul ont été utilisées par 72000 tâches de calcul. Ces tâches ont été distribuées sur 58 nœuds de grille répartis dans le monde. Ces nœuds de grille appartiennent à des centres de ressources regroupés par fédération en fonction de leur emplacement. La figure ci-dessous présente la distribution des tâches par pourcentage en fonction de la fédération.



La plus importante part des tâches a été envoyée au Royaume-Uni. Cela s'explique par la quantité de ressources qu'ils avaient mis à disposition de ce défi de production de données et par l'optimisation de l'utilisation des ressources réalisées d'une part par les services de grille et d'autre part par l'environnement de soumission de WISDOM. Au plus fort de la production, 1700 ordinateurs ont été utilisés simultanément. Le gain de temps réalisé par comparaison de la même production sur un unique ordinateur sur la même durée s'élève à 660.

Les résultats ont été stockés sur la grille en 2 exemplaires, pour une taille totale de 1 téraoctets. Les données, en cours d'analyse, sont accessibles facilement pour les partenaires WISDOM directement sur la grille.
La grille a montré, malgré des problèmes inhérents dû au système hétérogène et dynamique de l'infrastructure, qu'elle pouvait répondre à une demande importante et limitée dans le temps. Le docking s'est avéré également exploitable sur ce type d'outil.

V. Conclusion

Dans cet article, nous avons brièvement présenté les grilles informatiques et le potentiel qu'elles offrent dans le domaine de la santé. En particulier, nous avons recensés quels sont les avantages espérés d'une telle technologie pour combattre les maladies négligées. Et au-delà d'une vision, nous avons montrés que des premières applications de recherche de médicaments contre le paludisme étaient déployées à grande échelle dans un but de proposer de nouveaux inhibiteurs.
L'intégration de laboratoires de biologie expérimentale et de chimie assurera le suivi du criblage in silico réalisé sur grille dans le monde réel. Fermer le cercle entre les expériences du monde physique et les expériences virtuelles in silico est une perspective réaliste dans un futur proche.
A plus long terme, nous proposons le lancement d'une initiative internationale utilisant la technologie des grilles pour la découverte de nouveaux médicaments *in silico*, pour le développement de la collaboration entre les acteurs de la recherche et de l'intervention humanitaire et pour la mise en place d'une fédération de bases de données en zone endémique pour améliorer l'accès aux médicaments et la remontée de données épidémiologiques.

Remerciements

Les auteurs remercient particulièrement les projets EGEE, AuverGrid et Accamba ainsi que la Commission Européenne.

References

1. http://wisdom.eu-egee.fr/ or http://public.eu-egee.org/files/battles-malaria-grid-wisdom.pdf
2. http://www.africaathome.org/
3. http://www.swissbiogrid.com/project_and_proof.html
4. http://www.auvergrid.fr
5. Shin, JM., Cho, DH.; PDB-Ligand: a ligand database based on PDB for the automated and customized classification of ligand-binding structures. Nucleic Acids Res 2005 Jan 1 33 Database Issue:D238-41
6. Goto, S., Okuno, Y., Hattori, M., Nishioka, T. and Kanehisa, M.; LIGAND: database of chemical compounds and reactions in biological pathways. Nucleic Acids Res. 30, 402-404 (2002).
7. Kuntz, I.D., J.M. Blaney, S.J. Oatley, R. Langridge, and T.E. Ferrin. A Geometric Approach to Macromolecule-Ligand Interactions. Journal of Molecular Biology, 1982. 161:269-288.
8. Spencer, R. W: Highthroughput virtual screening of historic collections on the file size, biological targets, and file diversity. Biotechnol. Bioeng. 1998, 61, 6167.
9. Rajkumar Buyya, Kim Branson, Jon Giddy and David Abramson: The Virtual Laboratory. A Toolset to Enable Distributed Molecular Modeling for Drug Design on the WorldWide Grid (Unpublished)

10. Jochen Weisner, Regina Ortmann, Hasan Jomaa, Martin Schlitzer: Angew. New Antimalarial drugs. Chem. Int. Ed. 2003, 42, 5274529.
11. Francis, S. E., Sullivan, D. J. Jr., Goldberg, D. E: Hemoglobin metabolism in the malaria parasite plasmodium falciparum. Annu.Rev. Microbiol.1997, 51, 97123.
12. Coombs, G. H., Goldberg, D. E., Klemba, M., Berry, C., Kay, J., Mottram, J. C: Aspartic proteases of plasmodium falciparum and other protozoa as drug targets. Trends parasitol. 2001, 17, 532537.
13. Silva, A. M., Lee, A. Y., Gulnik, S. V., Majer, P., Collins, J., Bhat, T. N., Collins, P. J., Cachau, R. E., Luker, K. E., Gluzman, I. Y., Francis, S. E., Oksman, A., Goldberg, D. E., Erickson, J. W: Structure and inhibition of plasmepsin II, A haemoglobin degrading enzyme from Plasmodium falciparum. Proc. Natl. Acad. Sci. USA 1996, 93, 1003410039.
14. Morris, G. M., Goodsell, D. S., Halliday, R.S., Huey, R., Hart, W. E., Belew, R. K. and Olson, A. J: Automated Docking Using a Lamarckian Genetic Algorithm and Empirical Binding Free Energy Function. J. Computational Chemistry, 19,16391662. 1998.
15. M. Rarey, B. Kramer, T. Lengauer & G. Klebe: Predicting ReceptorLigand interactions by an incremental construction algorithm. J. Mol. Biol. 261,470489, 1996.
16. http://www.biosolveit.de
17. Irwin and Shoichet: J. Chem. Inf. Model. 2005;45(1),17782.
18. http://public.eu-egee.org
19. S. Campana et al. : Production experience on the LCG Computing Grid, presented at e-Science 2005 conference in Melbourne